Discriminative Bimodal Networks for Visual Localization and Detection with Natural Language Queries

University of Michigan

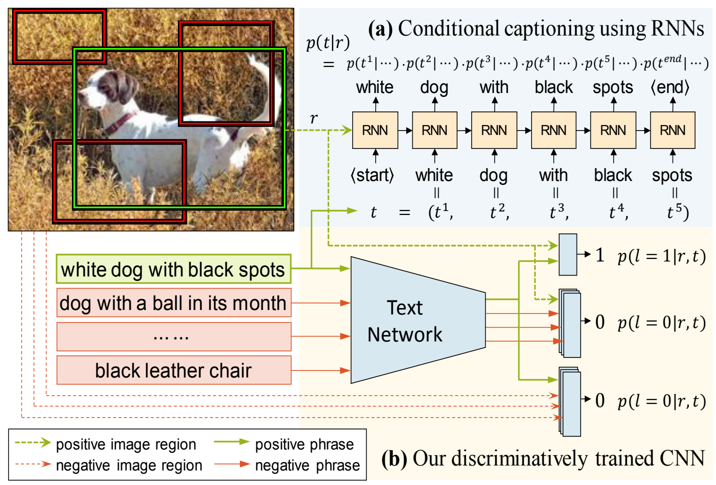
Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, Honglak Lee

Data, Code & Model: DBNet.link

1 Localization and Detection with Natural Language Queries

- **Tradition object detection:** person, cat, dog, car, motorbike, airplane, bed, television, …
- **Natural language query:** a group of bikers, a building with lots of windows, …

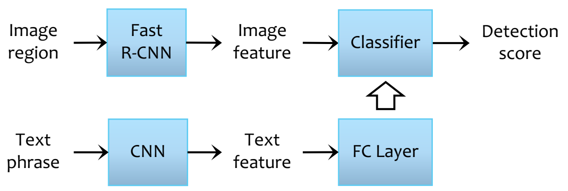
2 Discriminative Bimodal Networks (DBNet)



- **Typical previous works**
  - Caption generation model
  - Output in a huge language space
  - Mostly use only positive training samples (matched text and regions)
- **Our DBNet**
  - A fully discriminative model
  - Learning in a binary space is easier
  - Able to use all possible negative samples across the training set

3 Model Architectures

- Visual features: Fast R-CNN (VGGNet/ResNet)
- Text features: Character-level CNN
  - Replicating a phrase until reaching the input length of the CNN (256 characters)
  - Non-saturable activations: Leaky ReLU
  - Can be replaced by other text embedding



- The detection score of image region r is given by a linear classifier dynamically generated according to the text phrase t :

$$f(x, r, t; \Theta) = \mathbf{w}(t)^\top \phi_{\mathrm{rgn}}(x, r; \Theta_{\mathrm{rgn}}) + b(t).$$

where the classifier weights and bias are

$$\mathbf{w}(t) = \mathbf{A}_\mathbf{w}^\top \phi_{\mathrm{txt}}(t; \Theta_{\mathrm{txt}}), \;\; b(t) = \mathbf{a}_b^\top \phi_{\mathrm{txt}}(t; \Theta_{\mathrm{txt}}).$$

- An extra regularization term on the classifier parameter is important for the SGD stability:

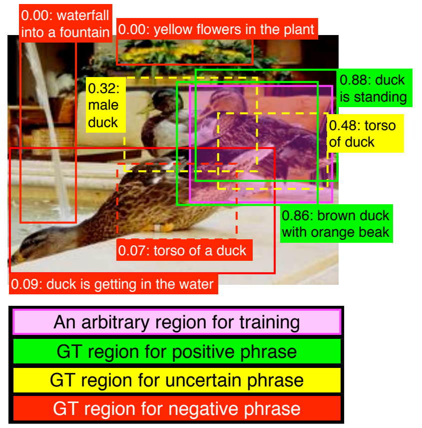
$$\Gamma_{\mathrm{dynamic}} = \|\mathbf{w}(t)\|_2^2 + |b(t)|^2$$

References

[DenseCap] J. Johnson, A. Karpathy, and L. Fei-Fei. "DenseCap: Fully convolutional localization networks for dense captioning". In *CVPR*, 2016.
[SCRC] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. "Natural language object retrieval". In *CVPR*, 2016.

CVPR 2017, July 21-26, Honolulu

4 Model Training: Labels, Objectives, and Optimization



- Any region-text pair (r, t) can be possibly used in training.
- r can be GT or proposed (EdgeBox), t can be an annotation on the same image as r or from the rest of the training set.
- **Spatial overlapping based training labels:**
  - **1:** r has large overlap with a ground truth region of t
  - **Uncertain:** (r, t) is not positive, and r has moderate overlap with a ground truth region of t. (excluded from training)

  Text similarity based uncertainty augmentation:
  If t' is similar to t, t' should also have uncertain label with r
  - **0:** otherwise (including all other text phrases).

- Given a region, the non-uncertain phrases are categorized into: positive phrases (pos), negative phrases from the annotations on the same image (neg), and negative phrases from the rest images (rest).
- Training loss is normalized separately for the three categories:

$$L_i = \lambda_{\mathrm{pos}} L_i^{\mathrm{pos}} + \lambda_{\mathrm{neg}} L_i^{\mathrm{neg}} + \lambda_{\mathrm{rest}} L_i^{\mathrm{rest}}$$

where we set $\lambda_{\mathrm{pos}} = \lambda_{\mathrm{neg}} + \lambda_{\mathrm{rest}}$, the sampling strategy in SGD determines λ_{neg} (larger) and λ_{rest} (smaller).

- Optimization: *(initialization)* Initializing visual pathway with pretrained faster R-CNN;
  (phase 1) training the text pathway from scratch and fix the visual pathway parameters;
  (phase 2) jointly finetuning the two pathway; *(phase 3) decrease the* learning rate.

5 Benchmarking Protocol

- Based on the Visual Genome dataset
  - Bounding boxes with text phrase descriptions
  - Additional spell checking and auto-correction
- **Localization:** find a known-to-be-existing entity
- **Detection:** localize all matched entities if exists any
  - Should have negative images (no matched entity)
  - Impractical to test all queries on each image

Scenario	# unique phrases
Training set	2,113,688
Test set	180,363
Test set (unseen on training)	119,315
Per image (level 0)	~43
Per image (level 1)	~92
Per image (level 2)	~775

- The first benchmark protocol for visual entity detection with language queries. 3 difficulty levels with increasing number of randomly chosen negative images per query.
  - Level 0: no negative image
  - Level 1: the same number as the positive image
  - Level 2: 5 times the number of positive images or 20 (whichever is larger) for each test phrase
- **Detection metric:** average precision (AP)
  - *mean AP (mAP):* averaging APs over all phrases; each phrase has its own decision threshold.
  - *global AP (gAP):* a single AP for any phrases; all phrases share the same decision threshold. (requires scores to be calibrated over phrases; for zero-shot settings)

6 Experiments

We release our implementation in both **Caffe+MATLAB** (original results) and **TensorFlow**.

- DBNet training is memory consuming. Our code perform subbatch partition for subnetworks to fit the GPU memory limit while keeping a large overall batch size for effective training.

a. Localization performance

Method	Accuracy/% for IoU@ 0.3	0.5	0.7	Median IoU
DenseCap	25.7	10.1	2.4	0.092
SCRC	27.8	11.0	2.5	0.115
DBNet w/o bias	36.3	22.4	9.4	0.124
DBNet	**38.3**	**23.7**	**9.9**	**0.152**
DBNet (ResNet)	**42.3**	**26.4**	**11.2**	**0.205**

b.1. Detection APs for IoU@0.3

Difficulty level:	0		1		2	
AP / %	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	36.2	1.8	22.9	1.0	4.1	0.1
SCRC	38.5	2.2	37.5	1.7	**29.3**	0.7
DBNet	**48.1**	**23.1**	**45.5**	**21.0**	26.7	**8.0**
DBNet (ResNet)	**51.1**	**24.2**	**48.3**	**22.2**	**29.7**	**9.0**

b.2. Detection APs for IoU@0.5

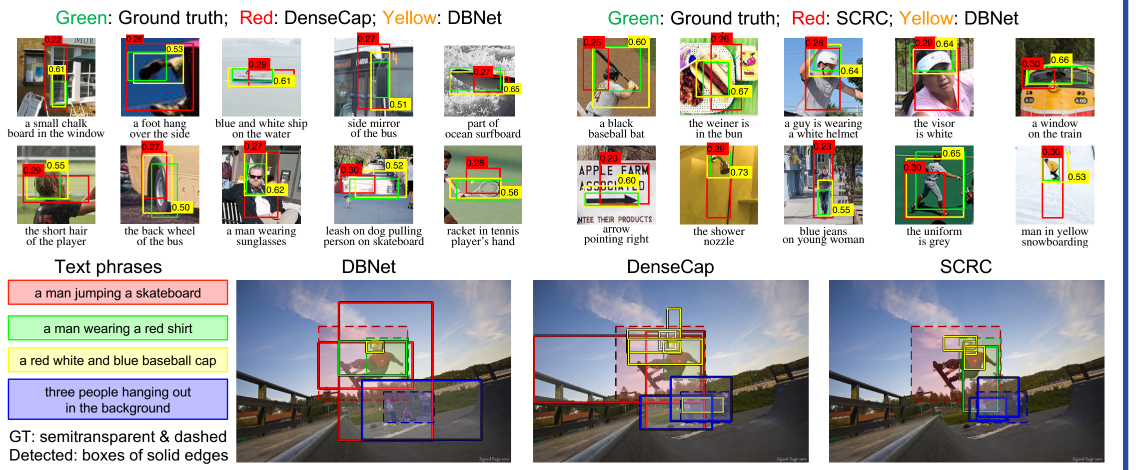
Difficulty level:	0		1		2	
AP / %	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	15.7	0.5	10.0	0.3	1.7	0.0
SCRC	16.5	0.5	16.3	0.4	12.8	0.2
DBNet	**30.0**	**10.8**	**28.8**	**9.9**	**17.7**	**3.9**
DBNet (ResNet)	**32.6**	**11.5**	**31.2**	**10.7**	**19.8**	**4.3**

b.3. Detection APs for IoU@0.7

Difficulty level:	0		1		2	
AP / %	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	3.4	0.0	2.1	0.0	0.3	0.0
SCRC	3.4	0.0	3.4	0.0	2.7	0.0
DBNet	**11.6**	**2.1**	**11.4**	**2.0**	**7.6**	**0.9**
DBNet (ResNet)	**12.9**	**2.2**	**12.6**	**2.1**	**8.5**	**0.9**

- DBNet outperforms captioning models.
- Dynamic bias term improves performance.
- DBNet shows higher per-phrase performance in term of mAP.
- DBNet's scores are better "calibrated" over different phrases in terms of gAP.
- DBNet shows more robustness to negative images and rare text phrases.

c. Quantitative comparison



e. Ablation study of DBNet's major components

Prune ambiguous phrases	Phrases from other images	Finetune visual pathway	Localization: Recall / % for IoU@ 0.3	0.5	0.7	Median IoU	Mean IoU	Detection (Level-1): mAP / % for IoU@ 0.3	0.5	0.7	gAP / % for IoU@ 0.3	0.5	0.7
No	No	No	30.6	17.5	7.8	0.066	0.211	35.5	22.0	8.6	8.3	3.1	0.4
Yes	No	No	34.5	21.2	9.0	0.113	0.237	39.0	24.6	9.7	15.5	7.4	1.6
Yes	Yes	No	34.7	21.1	8.8	0.119	0.238	41.3	25.6	10.0	17.2	7.9	1.6
Yes	Yes	Yes	**38.3**	**23.7**	**9.9**	**0.152**	**0.258**	**45.5**	**28.8**	**11.4**	**21.0**	**9.9**	**2.0**